Available online at www.sciencedirect.com

SCIENCE DIRECT®

Ecological Modelling 174 (2004) 161–173

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium)

Andy P. Dedecker*, Peter L.M. Goethals, Wim Gabriels, Niels De Pauw

Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent University, J. Plateaustraat 22, B-9000 Ghent, Belgium

Abstract

To meet the requirements of the EU Water Framework Directive, models are useful to predict communities in watercourses based on the abiotic characteristics of their aquatic environment. For that purpose back-propagation Artificial Neural Network (ANN) algorithms were used to induce predictive models on a dataset of the Zwalm river basin (Flanders, Belgium). This dataset consisted of 120 samples, collected over a 2-year period. Fifteen environmental variables were measured at each site, as well as the abundance of the aquatic macroinvertebrate taxa. Different neural networks were developed and optimized to obtain the best model configuration for the prediction of the habitat suitability of macroinvertebrate taxa. The best performing number of hidden layers and neurons and training algorithms have been searched for. The different options were theoretically and practically validated and assessed. The theoretical validation was based on cross-validation. For the practical validation, potential applications of the neural network models were analyzed, and the predictive performance of the models was assessed using ecological expert knowledge. The results indicate that the number of times a taxon was found in the whole river basin influences the performance measures and the architecture of the network. Based on the Cohen's kappa, it could be concluded that ANN models predicting the presence/absence of very rare taxa (e.g. *Aplexa*) or very common taxa (e.g. Tubificidae) were rather irrelevant, although their correctly classified instances (CCI) was high. Predicting the presence/absence of Asellidae (a moderately present taxon), the highest performances (CCI and Cohen's kappa) were found for the network model with two hidden layers each having 10 neurons. When calculation time was also taken into account, the network model with one hidden layer having 10 neurons could be preferred. Applying this network architecture, performances were only slightly worse, while calculation time was a lot shorter. One may also conclude that not all network models resulted in a relevant relation between a variable and a specific taxon. For Gammaridae for example, a rather small ANN structure gave a better idea of the impact of dissolved oxygen on its presence than a larger one. More reliable predictions and ecological interpretations for river ecosystem management would thus be possible provided the best configuration could be found.

Keywords: Optimization; Macroinvertebrates; Habitat preferences; Predictive modeling

1. Introduction

Human activities have severely deteriorated the Flemish river systems, and many functions such as drinking water supply, fishing, etc. are threatened. Because the restoration of these river systems entails drastic social (e.g. change in habits with regard to

* Corresponding author. Tel.: +32-9-264-37-08; fax: +32-9-264-41-99.
E-mail address: andy.dedecker@UGent.be (A.P. Dedecker).

0304-3800/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2004.01.003

water use and discharge, urban planning) and economical (e.g. investment in nature restoration, wastewater treatment system installation) consequences, the decisions should be taken with enough forethought. Ecosystem models could therefore act as interesting tools to support decision-making in river restoration management. In particular models that can predict the habitat requirements of organisms are of considerable importance to ensure that the planned actions have the desired effects on the aquatic ecosystems. It was shown that machine learning techniques such as Artificial Neural Networks (ANNs) basically mimic aspects of biological information processing for data modeling and could be useful in ecology (Recknagel, 2001). The prediction of aquatic communities by means of ANN models has recently been discussed by several authors (Brosse et al., 1999; Guégan et al., 1998; Hoang et al., 2001; Karul et al., 2000; Laë et al., 1999; Lee et al., 2003; Lek et al., 1996; Maier and Dandy, 1997; Maier and Dandy, 2001; Maier et al., 1998; Mastrorillo et al., 1997, 1998; Olden and Jackson, 2002; Park et al., 2001; Recknagel, 1997; Recknagel et al., 1997; Reyjol et al., 2001; Scardi, 2001; Scardi and Harding, 1999; Schleiter et al., 1999; Wagner et al., 2000; Wei et al., 2001; Wilson and Recknagel, 2001). It is stressed that the ANN architecture is generally highly problem dependent (Maier and Dandy, 2000). For this reason, it is necessary to develop and optimize the ANNs to obtain the best model configuration that gives lower error during training with minimal computing time. Traditionally, optimal network geometries have been found by trial and error (Brosse et al., 1999; Maier and Dandy, 2000). If predictions are made for different macroinvertebrate taxa, simultaneously another problem could emerge because the frequency of occurrence (the number of sites on which a taxon was found) could influence the ANN architecture. If the optimal ANN architecture could be found and reliable predictions would be possible, conclusions regarding ANN model design for practical use in ecological river management could be drawn.

The aim of this paper was to discuss the development and optimization of different neural network models to obtain the best model configuration for the prediction of macroinvertebrate taxa. Two taxa were selected: *Aplexa*, which is a very rare taxon in the Zwalm river basin (found at 4.2% of the sites), and Asellidae, which was present at 45.4% of the sites. The best performing network architecture and training algorithms were searched for. Finally, an ecological interpretation of the constructed models was made for Gammaridae.

2. Material and methods

2.1. Study sites and collected data

The Zwalm river basin which is part of the hydrographical basin of the Upper-Scheldt (Carchon and De Pauw, 1997) was selected as study area (Fig. 1). The basin has a total surface of 11,650 ha, the Zwalm river itself has a length of 22 km. The river has an irregular flow regime, ranging from 0.3 to 4.7 m^3/s. Although Flanders is in general a rather flat region, the Zwalm river basin is characterized by a number of differences in altitude, making it a quite unique ecosystem (Soresma, 2000). In the unpolluted headwaters a sensitive and vulnerable fauna is found (e.g. the bullhead (*Cottus gobio*) the brook lamprey (*Lampetra planeri*) and the mayfly Heptageniidae). Since 1999, the water quality in the Zwalm river basin has considerably improved due to investments in sewerage and wastewater treatment plants during the preceding years (VMM, 2000). Several parts of the river are however still polluted by untreated urban wastewater and by diffuse pollution originating from agricultural activities (Goethals and De Pauw, 2001). Numerous structural and morphological disturbances still exist (e.g. weirs for water quantity control, artificial embankments, etc.) (Carchon and De Pauw, 1997).

In total, 60 sites were selected in the Zwalm river basin at which physical and chemical samples were taken (Fig. 1). Observations regarding the structural characteristics were made. Each site was examined twice over a 2-year period (summer of 2000 and 2001). In this way, 120 sets of observations were available. Certain structural characteristics (meandering, substrate type, etc.) were visually monitored (Dedecker et al., 2002b). Flow velocity was determined by timing the transport of a float over a distance of 10 m. A number of flow velocity measurements at various places in the river (at the centre of the stream and at the bank side) were taken and the average figure was presented. Control

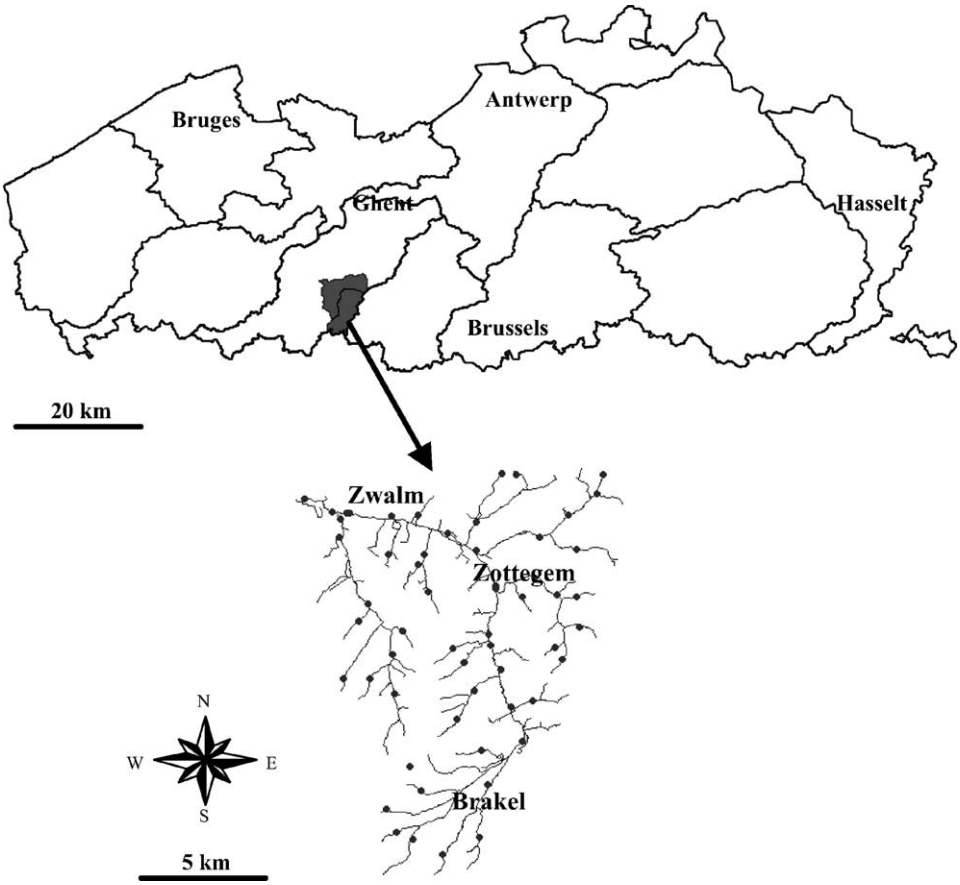

Fig. 1. Location of the Zwalm river basin in Flanders, Belgium. Position of the selected sampling sites in the Zwalm river basin.

measurements were done by means of a propeller. Field measurements were made for temperature and dissolved oxygen (OXI 330/SET), pH (Jenway 071) and conductivity (WTW LF 90). Suspended solids were measured spectrophotometrically in the laboratory (Dedecker et al., 2002b). Macroinvertebrates were collected by means of a standard handnet during 5-min kick sampling within a river stretch of 10 m (NBN, 1984) and by in situ exposure of artificial substrates (De Pauw et al., 1994). The objective of the sampling was to collect the most representative diversity of the macroinvertebrates at the examined site (De Pauw and Vanhooren, 1983). The structural characteristics and physico-chemical variables (Table 1) were used as inputs for the neural network models to predict the presence or absence (respectively represented by 1 and 0) of macroinvertebrate taxa in the headwaters and brooks of the Zwalm river basin.

2.2. Data processing

Because the input variables have very different orders of magnitude it is recommended to rescale the data. In this way, more reliable predictions can be made. The variables are rescaled to be included within the interval $[-1, 1]$ by using the following equation:

$$V_n = 2 \times \frac{V_o - V_{min}}{V_{max} - V_{min}} - 1 \quad (1)$$

in which V_o and V_n are respectively the old and new value of the variable for a sampling point, V_{min} and V_{max} are the minimum and maximum values of that variable in the original dataset. Also the targets are

Table 1
Abiotic input variables and units used in the ANN model

Variables	Units
Temperature	°C
pH	
Conductivity	μS/cm
Suspended solids	mg/l
Dissolved oxygen	mg/l
Water level	cm
Fraction of pebbles	% of river bed
Shade	%
Water plants	Present/absent
Width	cm
Flow velocity	m/s
Meandering	6 classes (1 = well developed to 6 = absent)
Hollow river beds	6 classes (1 = well developed to 6 = absent)
Deep/shallow variation	6 classes (1 = well developed to 6 = absent)
Artificial embankment structures	3 classes (0 = absent; 1 = moderate; 2 = intensive)

rescaled over the interval $[-1, 1]$ to adapt to the transfer function used (tangential sigmoid) in the output layer. In this way, the network will be trained to produce outputs in the range $[-1, 1]$. Afterwards, these outputs were converted back into the same units which were used for the original targets. The continuous network output is mapped to 0 and 1 using a threshold of 0.5.

2.3. Artificial Neural Networks

In this study, different neural network models were tested and optimized to obtain the best model configuration for the prediction of the habitat suitability of macroinvertebrate taxa. The modeling method was based on the principles of the backpropagation algorithm (Rumelhart et al., 1986). The construction of the ANN model was based on examples of data with known outputs. A backpropagation network typically comprises three types of neuron layers: an input layer, one or more hidden layers and an output layer each including one or more neurons. As shown in Fig. 2, nodes from one layer are connected to all nodes in the following layer, but no lateral connections within any layer, nor feed-back connections are possible. Fifteen input neuron are used, each representing an environmental variable. The output layer comprises one neuron, indicating the presence or absence of a macroinvertebrate taxon. With the exception of the input neurons, which only connect one input value with its associated weight values, the net input for each

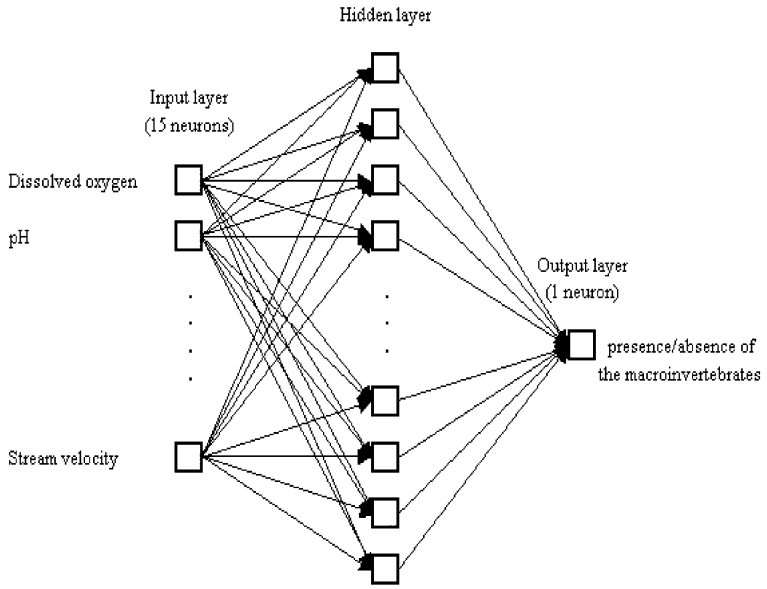

Fig. 2. Illustration of a three-layered neural network with one input layer, one hidden layer and one output layer.

neuron is the sum of all input values x_n, each multiplied by its weight w_{jn}, and a bias term z_j which may be considered as the weight from a supplementary input equalling one:

$$a_j = \sum w_{ji}x_i + z_j \quad (2)$$

The output value, y_j, can be calculated by feeding the net input into the transfer function of the neuron:

$$y_j = f(a_j) \quad (3)$$

Many transfer functions can be used. In this study, two types of sigmoid functions have been implemented: the logarithmic (for the hidden layer neurons) and tangential (for the output layer neurons) sigmoid transfer function. Layers of neurons with non-linear transfer functions allow the network to learn non-linear and linear relationships between input and output vectors. Thus they are ideally suited for the modeling of ecological data which are often known to be non-linear (Lek and Guégan, 1999).

Before training, the values of the weights and biases are initially set to small random numbers. Subsequently, a set of input/output vector pairs is presented to the network. For each input vector, the output vector is calculated by the neural network model, and an error term is calculated for the outputs of all hidden and output neurons, by comparing the calculated output vector and the actual output vector. Using this error term, the weights and biases are updated in order to decrease the error, so future outputs are more likely to be correct. This procedure is repeated until the errors become small enough or a predefined maximum number of iterations is reached. This iterative process is termed "training". After the training, the ANN can be validated using independent data.

In this study, two variations of the basic backpropagation algorithm have been compared to train the models: the gradient descent algorithm and the Levenberg–Marquardt algorithm (Hagan et al., 1996). The gradient descent algorithm updates the network weights and biases in the direction of the negative of the gradient. One iteration of this algorithm can be written as

$$x_{k+1} = x_k - \alpha_k g_k \quad (4)$$

in which x_k is a vector of current weights and biases, g_k is the current gradient, and α_k is the learning rate. The Levenberg–Marquardt algorithm is similar to the quasi-Newton method in which a simplified form of the Hessian matrix (second derivatives) is used. The Hessian matrix can be approximated as

$$H = J^{\mathrm{T}}J \quad (5)$$

and the gradient can be computed as

$$g = J^{\mathrm{T}}e \quad (6)$$

in which J is the Jacobian matrix which contains first derivatives of the network errors with respect to the weights and biases, and e is a vector of network errors. One iteration of this algorithm can be written as

$$x_{k+1} = x_k - [J^{\mathrm{T}}J + \mu I]^{-1}J^{\mathrm{T}}e \quad (7)$$

where μ is the learning rate and I the identity matrix (Hagan et al., 1996). During training the learning rate μ is incremented or decremented by a scale at weight updates. When μ is 0, this is just Newton's method, using the approximate Hessian matrix. When μ is large, this becomes gradient descent with a small step size. The Levenberg–Marquardt algorithm was reported to have the fastest convergence for neural networks that contain up to few hundred neurons (Karul et al., 2000).

The model validation was based on stratified 10-fold cross-validation (Witten and Frank, 2000). For 10-fold cross-validation the data are split into 10 folds or partitions. Each fold in turn is used for validation while the rest is used for training. That is, use nine-tenths for training and one-tenth for validation, and repeat the procedure 10 times so that in the end, every instance has been used exactly once for validation. To allow a reliable error estimate of the models, 10 stratified 10-fold cross-validation experiments were conducted. To compare the performances of the models trained with the gradient descent algorithm and the Levenberg–Marquardt algorithm and the models with different architectures, a paired *t*-test was done after checking for normality, to determine whether the mean of the set of samples was significantly greater or less. A paired *t*-test could be applied because the same splits were used to obtain a matched pair of results. For the two-tailed test, a significance level of 5% was used.

The models were evaluated on the basis of two performance measures: the percentage of correctly classified instances (CCI) and the Cohen's kappa (κ). For

Table 2
The confusion matrix as a basis for the performance measures with true positive values (TP), false positives (FP), false negatives (FN) and true negative values (TN)

Predicted	Actual	
	+	−
+	TP	FP
−	FN	TN

this one requires the derivation of matrices of confusion that identified true positive (TP), false positive (FP), false negative (FN) and true negative (TN) cases predicted by each model (Fielding and Bell, 1997). In that way, observed (actual) presence/absence patterns were tabulated against those predicted (Table 2).

The first performance measure that was calculated was the percentage of CCI:

$$\mathrm{CCI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \times 100 \tag{8}$$

Another performance measure that was calculated was the Cohen's kappa (Cohen, 1960). It is a simply derived statistic that measures the proportion of all possible cases of presence or absence that are predicted correctly by a model after accounting for chance predictions:

$$\kappa = \frac{(\mathrm{TP} + \mathrm{TN}) - [((\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP}) + (\mathrm{FP} + \mathrm{TN})(\mathrm{FN} + \mathrm{TN}))/n]}{n - [((\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP}) + (\mathrm{FP} + \mathrm{TN})(\mathrm{FN} + \mathrm{TN}))/n]} \tag{9}$$

To obtain the best model configuration for the prediction of the habitat suitability of *Aplexa* and Asellidae, two taxa with a different frequency of occurrence, two training algorithms were compared: the gradient descent algorithm and the Levenberg–Marquardt algorithm. For both training algorithms different network architectures were analyzed: five three-layered and four four-layered networks with respectively [2], [5], [10], [20], [25] and [5 5], [10 5], [10 10], [20 10] neurons in the hidden layer(s). The neural network models were implemented with the neural network extension of the software package MATLAB 5.3 for MS Windows™.

3. Results

3.1. Development and optimization of the ANN model configuration

The percentage of CCI and the Cohen's kappa for *Aplexa* (Mollusca) are shown in Figs. 3 and 4, respectively. Ten 10-fold cross-validations were conducted to obtain a reliable estimate for the performance measures. Also a 95% confidence interval of the average is shown. The CCI was high for both training algorithms. Based on the paired *t*-test (significance level of 5%), the CCI obtained with the gradient descent algorithm was not significantly different for the nine model architectures. The CCI was between 95.3 and 95.6%. The CCI obtained with the Levenberg–Marquardt algorithm was between 89.8 and 93.6%. When the architecture of the network models with the Levenberg–Marquardt algorithm becomes more complex, the CCI was significantly better, based on the paired *t*-test (significance level of 5%). A paired *t*-test was performed to compare the average CCI over ten 10-fold cross-validations of the gradient descent and the Levenberg–Marquardt algorithm. This statistical test could be used because the same cross-validation splits were used for both training algorithms. The results revealed a significant difference (significance level of 5%) between the CCI of the gradient descent and the Levenberg–Marquardt algorithm and that for all nine model architectures except for the architecture with 25 neurons in the hidden layer. For all the analyzed network architectures, the gradient descent algorithm resulted in a significant higher percentage of CCI. The second performance measure that was calculated was the Cohen's kappa, which accounts for the amount of chance predictions made by a model. Because the gradient descent algorithm predicted *Aplexa* absent at all sites, the Cohen's kappa remained 0 for all network architectures. The ANN models trained with the Levenberg–Marquardt algorithm had a Cohen's kappa between −0.02 and 0.01. Although a high CCI could be found for all network architectures, their Cohen's kappa indicated that this prediction success was merely based on chance. The Cohen's kappa of *Aplexa* was too low for both training algorithms, implying that these models cannot be considered relevant.

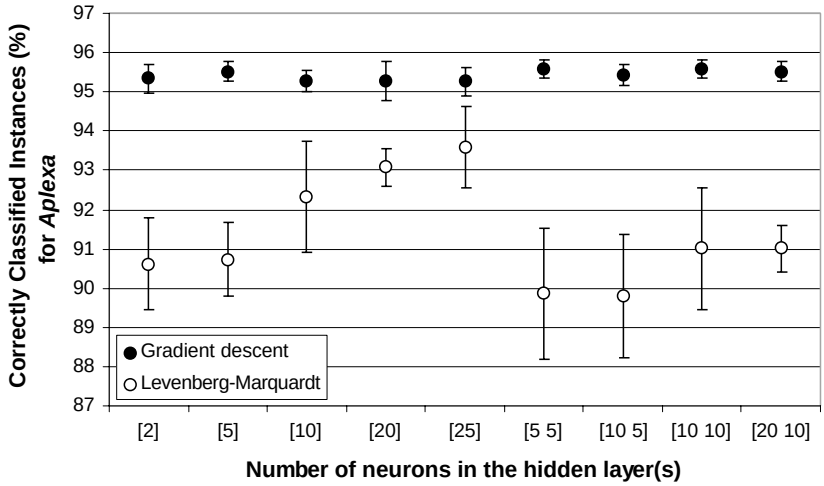

Fig. 3. Comparison of the percentage correctly classified instances for *Aplexa* (Mollusca) with the gradient descent and the Levenberg–Marquardt algorithm in different ANN architectures.

Contrary to *Aplexa*, Asellidae (Crustacea) are intermediately frequent in the Zwalm river basin. They were found at 45.4% of the sites. The CCI and the Cohen's kappa for Asellidae are shown in Figs. 5 and 6, respectively. The CCI was relatively high for both training algorithms. The CCI was between 73.9 and 76.6% for the gradient descent algorithm, and between 70.4 and 72.9% for the Levenberg–Marquardt algorithm. A Cohen's kappa between 0.45 and 0.51 was found for the gradient descent algorithm. The Cohen's kappa for the Levenberg–Marquardt algorithm was between 0.38 and 0.43. Based on Manel et al. (2001), a Cohen's kappa above 0.40 for presence/absence models is considered to indicate 'moderate' model performance while lower values indicates a low model performance. ANN models trained with the gradient descent algorithm outperformed ANN models trained with the Levenberg–Marquardt algorithm based on the CCI and the Cohen's kappa. The difference in performance was maximum 3.9%

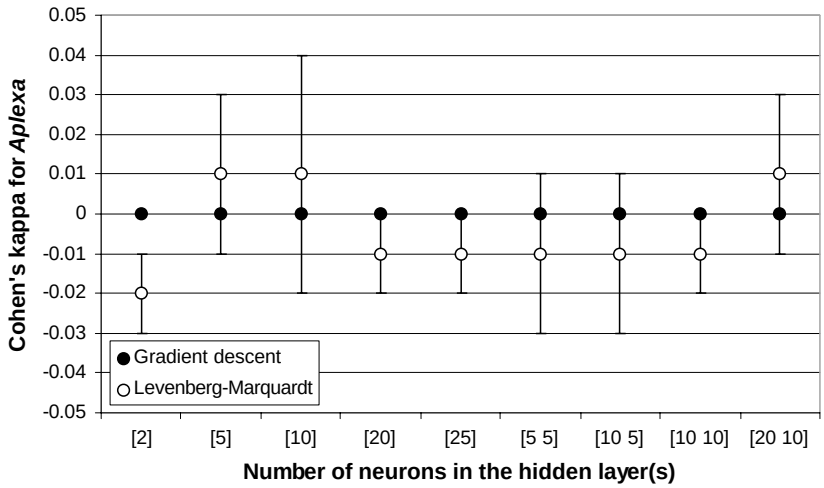

Fig. 4. Comparison of the Cohen's kappa for *Aplexa* (Mollusca) with the gradient descent and the Levenberg–Marquardt algorithm in different ANN architectures.

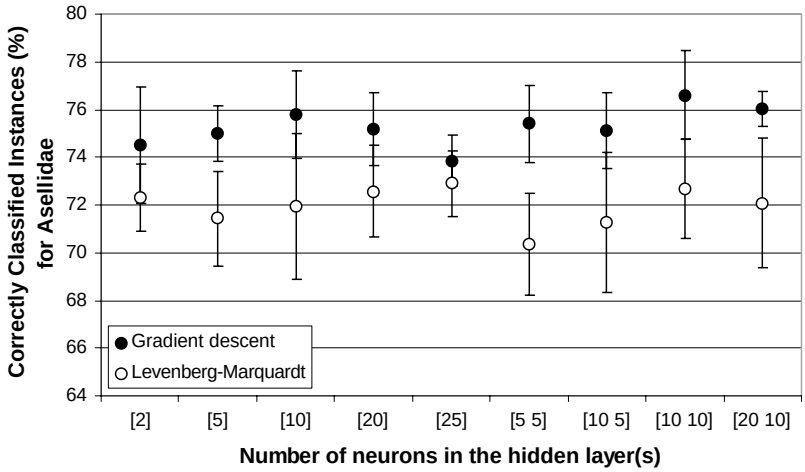

Fig. 5. Comparison of the percentage correctly classified instances for Asellidae (Crustacea) with the gradient descent and the Levenberg–Marquardt algorithm in different ANN architectures.

(network architecture [10 10] and [20 10]) based on the CCI and maximum 0.09 (network architecture [5 5]) based on the Cohen's kappa. Also a paired *t*-test was conducted to compare the average CCI and Cohen's kappa over ten 10-fold cross-validations of the gradient descent and the Levenberg–Marquardt algorithm. By applying this test, a significant difference (significance level of 5%) between the CCI of the gradient descent and the Levenberg–Marquardt algorithm was detected for all network architecture except for [2], [20] and [25]. Based on the Cohen's kappa, a significant difference (significance level of 5%) was found for the network architectures [5], [10], [5 5], [10 10] and [20 10]. Based on the CCI and the Cohen's kappa, the best performing network was the ANN model trained with the gradient descent algorithm with 10 neurons in both hidden layers. However, this ANN model was only significantly different (significance level of 5%) from the network [25].

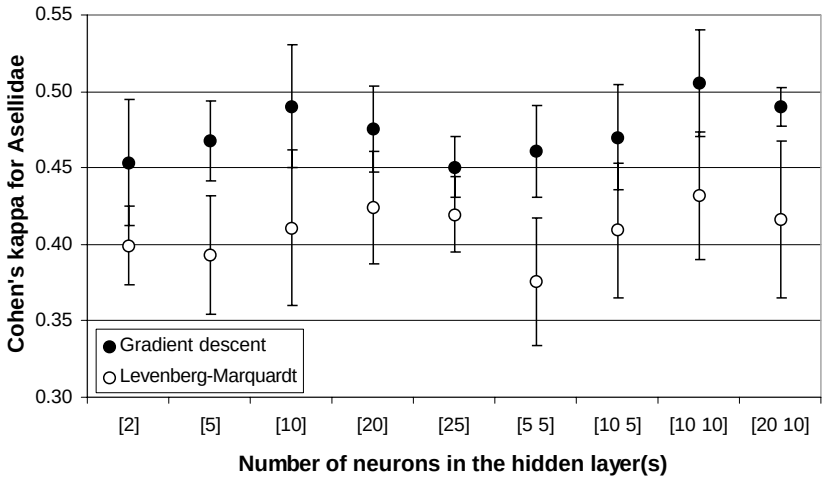

Fig. 6. Comparison of the Cohen's kappa for Asellidae (Crustacea) with the gradient descent and the Levenberg–Marquardt algorithm in different ANN architectures.

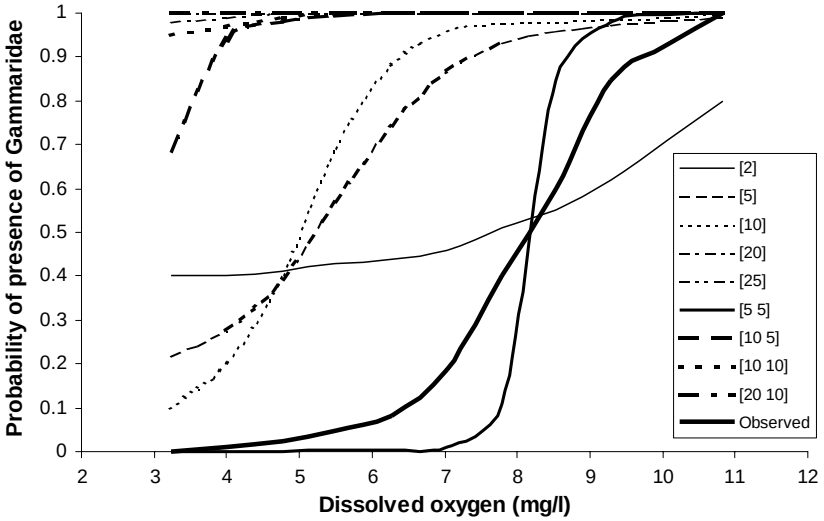

Fig. 7. The impact of dissolved oxygen on the probability of presence of Gammaridae (Crustacea). The predictions have been made for nine different network architectures. Also the cumulative curve of the observed presence of Gammaridae in the Zwalm river basin is presented.

3.2. Ecological interpretation of induced models

In ecology, it is useful to know the magnitude of impacts of a variable. Therefore, an experimental approach could be used to determine the response of the model to each of the input variables separately (Laë et al., 1999; Dedecker et al., 2002a). A range of variation of a single independent variable to the model is applied while the others are held constant. In this way, one is able to determine the impact of the variable on the presence or absence of a specific taxon. In order to limit the computation time, the number of points for each curve was limited to 11, delimiting 10 equal intervals over the variable. Different network architectures with the gradient descent algorithm were applied for each variable. Fig. 7 shows the effect of dissolved oxygen on Gammaridae (Crustacea). The networks [20], [25], [10 10] and [20 10] predicted Gammaridae as always present. The networks [2], [5], [10], [5 5] and [10 5] showed an increasing relation between dissolved oxygen and the predicted probability of presence of Gammaridae. Also the cumulative curve of the observed presence of Gammaridae in the Zwalm river basin is presented in Fig. 7.

4. Discussion

Maier and Dandy (2000) mentioned that network architecture is generally highly problem dependent. A number of systematic approaches for determining optimal network geometry have been proposed, including pruning and constructive algorithms. The basic thought of pruning algorithms is to start with a network that is large enough to capture the desired input–output relationship and to subsequently remove or disable unnecessary weights and/or neurons. Constructive algorithms approach the problem of optimizing the number of hidden layer neurons from the opposite direction to pruning algorithms. The smallest possible network is used at the start. Hidden layer neurons and connections are then added one at a time in an attempt to improve model performance. Traditionally, however, optimal network geometries have been found by trial and error (Brosse et al., 1999; Maier and Dandy, 2000). In this paper, trial and error is used to optimize the neural network architecture. Evaluating the predictive model performance frequently involves determining the percentage of sites for which presence or absence of organisms is correctly predicted (Manel et al., 2001). There is clear evidence though, that the CCI is influenced by

the frequency of occurrence of the organism being modeled (Dedecker et al., 2002a; Fielding and Bell, 1997; Manel et al., 1999). The problem with rare taxa is that there is little information to allow the neural network model to learn when these taxa are present. In this way the models tend to "learn" that very rare taxa are always absent. The same difficulty occurs with very common taxa. Here the models "learn" that very common taxa are always present. This problem is illustrated in the present study predicting the presence/absence of *Aplexa*. *Aplexa* was found at only 4.2% of the sites making it a rare taxon in the Zwalm river basin. The CCI was high for both training algorithms, namely between 95.3 and 95.6% for the gradient descent algorithm and between 89.8 and 93.6% for the Levenberg–Marquardt algorithm. As stressed by Manel et al. (2001) it is important to look at the predictions of the sites where the rare taxa are present and the common taxa are absent. Otherwise the evaluation of these models could be misleading. For this reason, it was decided to integrate an additional performance measure to assess the models, namely the Cohen's kappa (Cohen, 1960). The combination of CCI with Cohen's kappa, a measure of the proportion of all possible cases of presence or absence that are predicted correctly after accounting for chance effects, allowed a better interpretation of the predictive performance of the models (D'heygere et al., 2004). The ANN models trained with the gradient descent algorithm predicted *Aplexa* as always present, resulting in a Cohen's kappa of 0 for all network architectures. The models trained with the Levenberg–Marquardt algorithm were able to predict *Aplexa* as present in some cases. However, the network also predicted *Aplexa* as present at sites where they were not found, resulting in a Cohen's kappa between −0.02 and 0.01. Based on the Cohen's kappa, it could be concluded that the produced models for *Aplexa* were irrelevant, although their CCI was high. For the very common taxa, similar conclusions could be drawn. Contrary to *Aplexa*, good model performances were obtained for Asellidae which were found at 45.4% of the sites. The CCI was relatively high for both training algorithms. However, based on the CCI the ANN models trained with the gradient descent algorithm outperformed the models trained with the Levenberg–Marquardt algorithm for all network architectures. Also the Cohen's kappa was higher for all network architectures trained with the gradient descent algorithm. Based on Manel et al. (2001), a Cohen's kappa above 0.40 for presence/absence models is considered to indicate 'moderate' model performance while lower values indicate a low model performance. In this way, the Cohen's kappas obtained with the gradient descent algorithm can be classified as 'moderate' performances, while some of the Cohen's kappas obtained with the Levenberg–Marquardt algorithm indicate a low model performance. Based on these two performance measures, a neural network model trained with the gradient descent algorithm is preferred. Another drawback of the Levenberg–Marquardt algorithm in comparison with the gradient descent algorithm was the long calculation time for the more complex network architectures by its demand for memory to operate with large Jacobians and a necessity of inverting large matrices. The rank of matrices to be inverted is equal to the number of weights in the system. Such large matrices must be inverted at each iteration step and this results in large computation time. The highest performances (CCI and Cohen's kappa) were found for the network model with two hidden layers each having 10 neurons. In this way, this is the most appropriate model to predict the presence/absence of Asellidae. Although, when calculation time is also taken into account, the network model with one hidden layer having 10 neurons could be preferred. Applying this network architecture, performances were only slightly worse, while calculation time was a lot shorter. As demonstrated, the predictive ability of a given network architecture and training algorithm depends on the frequency of presence of a given macroinvertebrate taxon. If in another case, the frequency of a taxon is unknown a priori, the choice of a suitable model can be based on data from other studies which discuss similar systems. These comparable systems could give an indication whether a macroinvertebrate taxon is expected to be very rare, very common or rather moderately present. If these studies point out, a macroinvertebrate taxon is expected to be moderately present, predictions could be based on ANN models, because good results could be obtained as shown in this study. On the other hand, if a taxon is expected to be very rare or very common, predictions could be based on expert knowledge based Fuzzy Logic models, in which external expert knowledge can be incorporated, because ANN predictions seems

to be rather irrelevant in these cases based on this study.

Further optimization of the ANN models can be obtained by the selection of more appropriate input variables using, e.g. genetic algorithms (D'heygere et al., 2002, 2004). The variables that are not selected can be seen as irrelevant for a particular taxon (Witten and Frank, 2000). In ANN models, the irrelevant information is also sent through the nodes and can as such slightly alter the connection weights and affect the overall performance of ANNs (D'heygere et al., 2004). In this study, a set of parameters including learning rate, momentum and threshold value is held constant. In the future, genetic algorithms will also be used to automatically calibrate these parameters of the network (D'heygere et al., 2004).

In many studies ANN models have been shown to reveal superior predictive power compared to traditional approaches, e.g. multiple regression (Lek et al., 1996). Although, a disadvantage of ANN models in comparison with conventional models is their lack of explanations regarding the relative importance of each independent variable considered. In this way, ANN models have been labeled a 'black box'. This lack of illustrative power is a major concern to ecologists since the interpretation of statistical models is desirable for gaining knowledge of the causal relationships driving ecological phenomena (Olden and Jackson, 2002). Olden and Jackson (2002) describe a number of methods for understanding the mechanics of ANN models. They propose a randomization test for ANN models, which provides a statistical pruning technique for eliminating null connection weights that minimally influence the predicted output, as well as provides a selection method for identifying independent variables that significantly contribute to network predictions. Özesmi and Özesmi (1999) proposed the neural interpretation diagram (NID) for providing a visual interpretation of the connection weights among neurons, where the relative magnitude of each connection weight is represented by line thickness and line shading represents the direction of the weight. Garson (1991) proposed a method for partitioning the neural network connection weights in order to determine the relative importance of each input variable in the network. A number of investigators have used sensitivity analysis (Dedecker et al., 2002a; Guégan et al., 1998; Laë et al., 1999; Lek et al., 1996; Mastrorillo et al., 1998), varying the input variable across its entire range while holding all other input variables constant, so that the individual contributions of each variable are assessed. In this work a sensitivity analysis has been performed. As mentioned in the literature, Gammaridae prefer relatively high levels of dissolved oxygen. This relationship can also be derived from the induced models (Fig. 7). However, not all the networks gave this relationship. The most complex networks, networks [20], [25], [10 10] and [20 10], predicted Gammaridae as always present, which is ecologically inappropriate. The presence of Gammaridae in the Zwalm river basin was analyzed composing the cumulative curve of the observed values. Comparing this curve with the predicted probability of presence of Gammaridae, the best approach is given by the network with two hidden layers [5 5]. However, when the best network model was used, sensitivity analysis provided useful insight in the habitat preference of that taxon, which means important information for river ecosystem management. Laë et al. (1999) for example illustrated the influence of six independent environmental variables on the fish yield in the ANN modeling. For most variables the authors found ecologically relevant relations. This is in contrast to this research where the relations between some variables and the presence/absence of the macroinvertebrate taxa were difficult to interpret, although the predictive performance of the ANN models was in general good.

5. Conclusions

Artificial Neural Network models are efficient tools to predict the occurrence of macroinvertebrate taxa based on the abiotic characteristics of their aquatic environment. Several authors proved that ANN models are good alternatives for traditional approaches such as multiple regression (Lek et al., 1996). As mentioned before, the network structure to be used is very problem dependent. The results of this research indicate also that the frequency of occurrence of a taxon in the whole river basin influences the performance measures and the architecture of the network. Based on the Cohen's kappa, it could be concluded that ANN models predicting the presence/absence of very rare taxa (e.g. *Aplexa*) or very common taxa (e.g. Tubificidae) are rather irrelevant, although their CCI is high.

Predicting the presence/absence of Asellidae (a moderately present taxon), the highest performances (CCI and Cohen's kappa) were found for the network model with two hidden layers each having 10 neurons. When calculation time is also taken into account, the network model with one hidden layer having 10 neurons can be preferred. Applying this network architecture, performances are only slightly worse, while calculation time is a lot shorter. One might also conclude that not all network models are capable of finding a relevant relation between a variable and a specific taxon. For the Gammaridae, for example, a rather small network structure gave a better idea of the impact of dissolved oxygen than a larger one. The challenge will be to build the best model configuration, if more reliable predictions are to be expected. This is essential for a correct ecological interpretation, needed for ecosystem management.

Acknowledgements

The first author is a recipient of a grant of the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT).

References

Brosse, S., Guégan, J.F., Tourenq, J.N., Lek, S., 1999. The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecol. Model. 120 (2/3), 299–311.

Carchon, P., De Pauw, N., 1997. Development of a methodology for the assessment of surface waters. Study by order of the Flemish Environment Agency (VMM), Ghent University, Laboratory of Environmental Toxicology and Aquatic Ecology, Ghent, Belgium, 55 pp. (in Dutch).

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.

Dedecker, A.P., Goethals, P.L.M., De Pauw, N., 2002a. Comparison of artificial neural network (ANN) model development methods for prediction of macroinvertebrate communities in the Zwalm river basin in Flanders, Belgium. TheScientificWorldJOURNAL 2, 96–104.

Dedecker, A., Goethals, P.L.M., Gabriels, W., De Pauw, N., 2002b. Optimisation of Artificial Neural Network (ANN) model design for prediction of macroinvertebrate communities in the Zwalm river basin (Flanders, Belgium). In: Rizzoli, A.E., Jakeman, A.J. (Eds.), Integrated Assessment and Decision Support Proceedings of the 1st Biennial Meeting of the International Environmental Modelling and Software Society, vol. 2. SEA, Como, pp. 142–147.

De Pauw, N., Lambert, V., Van Kenhove, A., Bij De Vaate, A., 1994. Performance of two artificial substrate samplers for macroinvertebrates in biological monitoring of large and deep rivers and canals in Belgium and The Netherlands. Environ. Monit. Assess. 30, 25–47.

De Pauw, N., Vanhooren, G., 1983. Method for biological assessment of watercourses in Belgium. Hydrobiologia 100, 153–168.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2002. Use of genetic algorithms to select input variables in Artificial Neural Network models for the prediction of benthic macroinvertebrates. In: Rizzoli, A.E., Jakeman, A.J. (Eds.), Integrated Assessment and Decision Support Proceedings of the 1st Biennial Meeting of the International Environmental Modelling and Software Society, vol. 2. SEA, Como, pp. 136–141.

D'heygere, T., Goethals, P.L.M., De Pauw, N., 2004. Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. Ecol. Model. in press.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49.

Garson, G.D., 1991. Interpreting neural-network connection weights. Artif. Intell. Expert 6, 47–51.

Goethals, P.L.M., De Pauw, N., 2001. Development of a concept for integrated ecological river assessment in Flanders, Belgium. J. Limnol. 60 (1), 7–16.

Guégan, J.F., Lek, S., Oberdorff, T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391, 382–384.

Hagan, M.T., Demuth, H.B., Beale, M., 1996. Neural Network Design. PWS Publishing Company, Boston, 712 pp.

Hoang, H., Recknagel, F., Marshall, J., Choy, S., 2001. Predictive modelling of macroinvertebrate assemblages for stream habitat assessments in Queensland (Australia). Ecol. Model. 195, 195–206.

Karul, C., Soyupak, S., Cilesiz, A.F., Akbay, N., Germen, E., 2000. Case studies on the use of neural networks in eutrophication modeling. Ecol. Model. 134, 145–152.

Laë, R., Lek, S., Moreau, J., 1999. Predicting fish yield of African lakes using neural networks. Ecol. Model. 120, 325–335.

Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modelling of coastal algal blooms. Ecol. Model. 159, 179–201.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90, 39–52.

Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecol. Model. 120, 65–73.

Maier, H.R., Dandy, G.C., 1997. Modelling cyanobacteria (blue–green algae) in the River Murray using artificial neural networks. Math. Comput. Simul. 43 (3–6), 377–386.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Model. Software 15, 101–124.

Maier, H.R., Dandy, G.C., 2001. Neural network based modelling of environmental variables: a systematic approach. Math. Comput. Model. 33, 669–682.

Maier, H.R., Dandy, G.C., Burch, M.D., 1998. Use of artificial neural networks for modelling cyanobacterial *Anabaena* spp. in the River Murray, South Australia. Ecol. Model. 105 (2/3), 257–272.

Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J., 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. J. Appl. Ecol. 36, 734–747.

Manel, S., Wililiams, H.C., Ormerod, S.J., 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38, 921–931.

Mastrorillo, S., Dauba, F., Oberdorff, T., Guégan, J.F., Lek, S., 1998. Predicting local fish species richness in the Garonne river basin. Ecology 321, 423–428.

Mastrorillo, S., Lek, S., Dauba, F., Belaud, A., 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshwater Biol. 38, 237–246.

NBN, 1984. Norme Belge T 92-402. Biological water quality: determination of the biotic index based on aquatic macro-invertebrates. Institut Belge de Normalisation (IBN) (in Dutch and French).

Olden, J.D., Jackson, D.A., 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecol. Model. 154, 135–150.

Özesmi, S.L., Özesmi, U., 1999. An artificial neural network approach to spatial habitat modeling with interspecific interaction. Ecol. Model. 116, 15–31.

Park, Y.S., Kwak, I.S., Chon, T.S., Kim, J.K., Jorgensen, S.E., 2001. Implementation of artificial neural networks in patterning and prediction of exergy in response to temporal dynamics of benthic macroinvertebrate communities in streams. Ecol. Model. 146, 143–157.

Recknagel, F., 1997. ANNA—artificial neural network model for predicting species abundance and succession of blue–green algae. Hydrobiologia 349, 47–57.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K.I., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Model. 96 (1–3), 11–28.

Recknagel, F., 2001. Applications of machine learning to ecological modeling. Ecol. Model. 146, 303–310.

Reyjol, Y., Lim, P., Belaud, A., Lek, S., 2001. Modelling of microhabitat used by fish in natural and regulated flows in the river Garonne (France). Ecol. Model. 146, 131–142.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagation errors. Nature 323, 533–536.

Scardi, M., 2001. Advances in neural network modeling of phytoplankton primary production. Ecol. Model. 146, 33–45.

Scardi, M., Harding, L.W., 1999. Developing an empirical model of phytoplankton primary production: a neural network case study. Ecol. Model. 120 (2/3), 213–223.

Schleiter, I.M., Borchardt, D., Wagner, R., Dapper, T., Schmidt, K.D., Schmidt, H.H., Werner, H., 1999. Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. Ecol. Model. 120 (2/3), 271–286.

Soresma, 2000. Environmental impact assessment report on the development of fish migration channels and natural overflow systems in the Zwalm River basin, Soresma advies- en ingenieursbureau, Antwerp (in Dutch).

VMM, 2000. Water Quality—Water Discharges 1999. Flemish Environmental Agency, VMM, Erembodegem (in Dutch).

Wagner, R., Dapper, T., Schmidt, H.H., 2000. The influence of environmental variables on the abundance of aquatic insects: a comparison of ordination and artificial neural networks. Hydrobiologia 422/423, 143–152.

Wei, B., Sugiura, N., Maekawa, T., 2001. Use of artificial neural network in the prediction of algal blooms. Water Res. 35 (8), 2022–2028.

Wilson, H., Recknagel, F., 2001. Towards a generic artifical neural network model for dynamic predictions of algal abundance in freshwater lakes. Ecol. Model. 146, 69–84.

Witten, I.H., Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, San Francisco, 369 pp.